Queen Margaret University ultrasound, audio and video multichannel recording facility (2008-2016)

Alan A. Wrench & James M. Scobbie

Working Paper WP-24

June 2016

Queen Margaret University
CLINICAL AUDIOLOGY, SPEECH AND LANGUAGE RESEARCH CENTRE

Update and Sourcing Information June 2016

This paper is available online in pdf format

- 2016 onwards at www.qmu.ac.uk/casl
  and http://eresearch.qmu.ac.uk/id/eprint/4367

Author Contact details:

- 2016 onwards at awrench@articulateinstruments.com

Subsequent publication & presentation details:

- This is version 1 of this working paper – if amendments or clarifications are incorporated, or if the lifetime of the facility is extended, revised versions may be created and put online.
- We do not intend to publish this report except online at QMU.
- We hope to add some audio-visual support materials to augment this written report.

© The Authors 2016

This series consists of unpublished "working" papers. They are not final versions and may be superseded by publication in journal or book form, which should be cited in preference.

All rights remain with the author(s) at this stage, and circulation of a work in progress in this series does not prejudice its later publication.

Comments to authors are welcome.

CASL Research Centre Working Paper WP-24 (2016)
Series Editor: James M Scobbie

Queen Margaret University
ultrasound, audio and video multichannel recording facility (2008-2016)

Alan A Wrench[*†] and James M Scobbie[*]

* Queen Margaret University, † Articulate Instruments Ltd.

Abstract

This working paper describes some of the technical characteristics of the QMU speech recording facility, used to record high speed ultrasound, video, audio and other channels. It has been used by a number of projects at Queen Margaret University, in the Clinical Audiology, Speech and Language Research Centre since 2008, and some technical information has been made available in a range of publications, manuals and student reports. Here we collate in one place the background information relevant to issues of spatial resolution, time resolution, temporal synchronisation, edge detection and confidence, and comment on their general relevance for ultrasound-based speech research at QMU and in other laboratories.

1 Introduction

The QMU high-speed ultrasound facility is centred around an Ultrasonix RP which was acquired in 2006. Between 2006 and 2008 bespoke software was developed to remotely control the RP from a control PC using the Ulterius software development kit provided by Ultrasonix.[1]

The RP runs scanning software called "Exam" and the QMU facility runs version 3.17 of Exam. This is an old revision, but provides the facility for quickly transferring raw pre-scan digitised data from the RP scanner's internal memory to the remote PC in real

[1] Hardware in general has been funded by research grants and projects, e.g. the high-speed Ultrasonix RP machine by SHEFC / SRIF-3 (Strategic Research Infrastructure) Grant 2006-2008 (£121,369). Initially some core software functionality to handle ultrasound data had been commissioned from Articulate Instruments Ltd, implemented as add-on modules for the company's EPG software (Articulate Assistant™), importing video output signals from an ultrasound scanner, but most of the facility software described here for the high-speed ultrasonix system was subsequently developed by Articulate Instruments independently and commercialised independently. QMU has the same financial relationship to Articulate Instruments Ltd as other customers, thought the two entities have a close working relationship and shared strategic interests.

time. Newer revisions of Exam (> 5.7) always transfer the radio frequency data as well as the required envelope detected raw data, increasing the transfer time to an unacceptably long duration.

The bespoke software was called Articulate Assistant Advanced™ (AAA for short), and has a different underlying architecture. AAA permits a number of data channels to be started or stopped with the click of a single button and these channels are synchronised within the software. Once captured, data can be analysed using the same software environment. AAA includes a large number of data analysis options including annotation for segmental or gestural labelling, automated tongue surface contour tracking, dynamic kinematic measures for distance, area or velocity measures (etc), a polar statistical contour comparison measure, formant tracking, data export, file export and many more.[2]

Though initially developed for this specific RP machine, AAA is now more general and is used with a number of set-ups. Here we describe the specific implementation at QMU, though we make comments in passing about other similar set-ups.

2 Synchronisation

AAA stores every data stream in a database structure. Every channel other than audio has data samples tagged with timestamps relative to the audio channel. By using an Adlink 2213 A/D card configured to provide 8 differential input analogue input channels with independent gain controls and frequency response down to DC, all the analogue channels are implicitly synchronised. In the typical QMU setup this leaves ultrasonix, video, and EPG data to be synchronised explicitly.

Ultrasonix RP is unusual (along with other ultrasonix research models plus comparable systems from other manufacturers, such as the EchoB and Sonospeech systems provided by Articulate instruments) in that it generates a 3 nanosecond pulse the instant that each beamformed pulse is received in full. The last pulse in a B-mode scan that is used to create a single 2D image frame has a special use. This last pulse, which is generated after the frame is complete, can be transmitted on its own. It is this TTL level frame sync pulse that is used to precisely align the ultrasound with the audio. This TTL signal is processed and connected to one of the 8 analogue differential input channels, and this ensures it is time-locked with the microphone (audio) input. Since 3 nanoseconds is too short a pulse duration to show up on a standard analogue A/D channel recording at 22 or even 44kHz, a purpose-built unit (also commercially available) called a PStretch stretches out the pulse duration by a factor of 1000 from 3 nanoseconds to 3ms. Importantly though, the leading edge of this 3ms pulse is at the same instant as the leading edge of the 3 nanosecond pulse: it is this edge that is detected by the AAA software for synchronisation.

[2] Manuals, tutorial videos, and a free sample of data which can be explored with a limited free version of the software (unable to load or save other data) can be found at www.articulateinstruments.com

Immediately before recording, scanning is paused. To start a recordings, the audio recording is initiated first, and then the ultrasound is initiated. In this way, the time of first pulse leading edge on the recorded sync signal from ultrasonix can be measured to the nearest audio sample and the first frame of ultrasound data is tagged with this time. This is repeated for every subsequent frame. We are guaranteed no dropped frames as the data is recorded into the RP cineloop then transferred over local Ethernet using the robust TCP-IP protocol that is standard for lossless network data transfer.

EPG (electropalatography) data, if captured, is synchronised in an almost identical way to the Ultrasonix data. A third A/D channel is used to record a similar sync signal generated by the WinEPG system and AAA detects the pulses and tags the EPG frames accordingly.

Video synchronisation is performed slightly differently. A pulse, generated by the WinEPG system is used to trigger a bespoke Articulate Instruments "BrightUp" unit. This unit intercepts the analogue NTSC video signal output from a camera or other video device and merges (effectively, superimposes) a white square in the top left corner of 4 de-interlaced frames immediately following the trigger pulse from the WinEPG system.

The AAA software provides a dialogue to allow the NTSC video channel to be automatically batch synchronized for a whole set of recordings. The series of video images are equally time-stepped to a frame rate (see below) and de-interlaced (i.e. each frame is split into two, using either the odd or the even horizontally-organised lines from which the image is composed). De-interlacing provides an image stream at double the number of frames per second at half the vertical resolution, i.e. 640x240 pixels rather than 640x480. The odd and even deinterlaced images each have a unique timepoint and content, as an NTSC camera process records each of them independently, in two separate passes of the image *before* interlacing them together, the original purpose being to reduce visible image flicker. Thus all NTSC videos have rediced the underlying temporal resolution by a factor of two by interlacing the scans. The de-interlacing operation simply reverses this operation. The images in each frame are then stretched vertically to restore the original 4:3 aspect ratio providing images that appear to be 640x480 (i.e. each horizontal line of pixels is repeated).

The batch synchronisation process finally detects and aligns the first video frame containing the white square with the pulse from the WinEPG system that triggered it. The pulse is independently recorded as part of the EPG synchronisation process on an audio channel. This synchonisation links the audio and video channels at the time the pulse was created, near the start of the recording, but does not necessarily ensure that the video remains in sync throughout a recording. For video to remain in sync the video frame rate must be accurately estimated with respect to the audio channel. Typically it is assumed that the camera frame rate is close to the NTSC broadcast standard of 29.97Hz (59.94Hz de-interlaced), and the hardware can be checked against a reference signal. It is further assumed the nominal sample rate selected on the Adlink

A/D card is accurate. On our system, over the course of a typical 6-15 second recording, no loss of synchronisation has been observed.

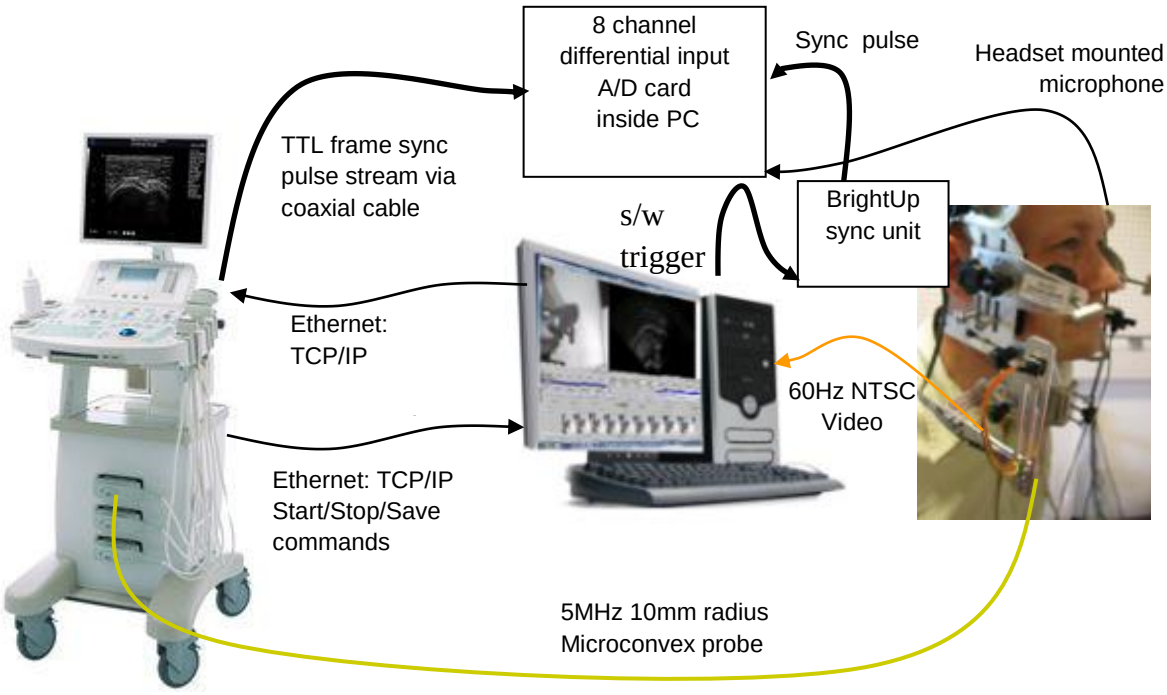


Figure 1. Connection diagram for the Ultrasonix, video and audio channels.

3 Spatial and temporal resolution of the Ultrasonix data

Ultrasonix RP (along with other ultrasonix research models and EchoB and Sonospeech systems provided by Articulate Instruments) is unusual for ultrasound machines used in phonetics research in that it provides direct access to the envelope detected and digitised beam-formed scanline data, rather than just to NTSC video output or to the contents of a limited cineloop buffer. The RP instrument provides a vector of 412 8-bit values for the echo detection results of each beam emanating from the transducer array. The number of such RP beams making up each frame is flexible, and can be selected using a parameter in the AAA software to control the spatiotemporal resolution of the data. It is important to understand that these beams do not correspond to an ultrasound emission emanating from a single crystal in the transducer array, but rather a beam is a complex cluster of emissions "steered" by firing several neighbouring piezo elements with carefully timed delays so as to form an ultrasonic wavefront with an interference pattern that is constructive in the intended direction of travel. Such a beam is not linear vector – it has a radial width and a thickness, both of which have implications for spatial resolution, and are considered later in this paper.

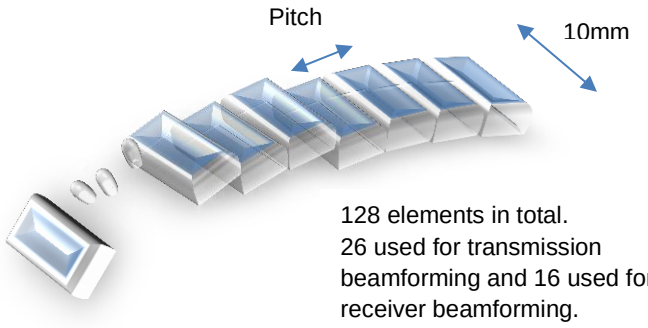


Figure 2. Makeup of the 128 element 10mm radius 5MHz microconvex ultrasonix transducer.

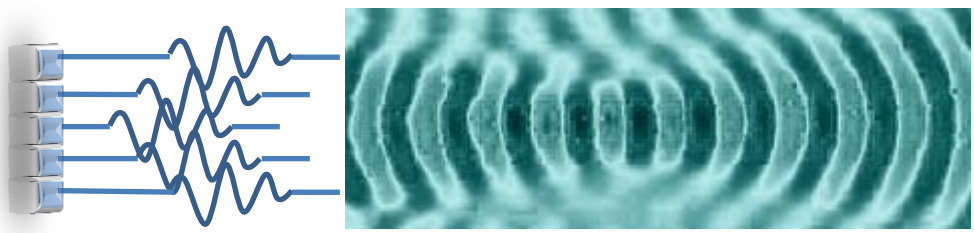


Figure 3. A single ultrasonic "beam" or "scanline" generated by firing several array elements to form a constructive interference pattern.

Control over the number of beams per frame gives control of the frame rate independent of the depth and field of view (FOV) settings. It is these three parameters plus the speed of sound through tissue that dictate the frame rate. For a typical recording setup the following parameters are chosen.

Table 1. Typical Ultrasonix parameter settings

- Number of beams/scanlines per 100% FOV = 69.
- FOV = 90% of 156 degree maximum, resulting in 63 beams over 135° FOV
- Maximum depth from probe surface = 80mm

- Speed of sound = 1540m/s

The depth setting determines the amount of time that the machine waits for echoes to return before sending out the next beam. So for a depth setting of 80mm the system must wait 2 x 0.08/1540 seconds = 104 µs. Multiplying this time by the number of beams making up a frame (63 in this case) defines the absolute limit on frame rate, which in this example is 152Hz. However there is a small processing overhead for each beam, and for each frame, so these parameters in fact result in an operational frame rate of 121Hz (to the nearest integer).

It is worth noting that this is a physical limit which applies to nearly all diagnostic ultrasound systems. Consequently any other manufacturer, make or model of ultrasound system operating at a frame rate of 121Hz at a depth setting of 80mm will also comprise approximately 63 beams, and the absolute upper limit would be 74 beams if there were no processing overheads. This information is not typically made available to users of other systems, resulting in misapprehensions about the resolution of systems used in other laboratories. In fact, the sparsity of the underlying beams can be easily seen in the angular smearing (lack of circumferential resolution) which is clearly visible in most ultrasound images. An example of this for a GE Logiq running at 124fps is shown in figure 4.

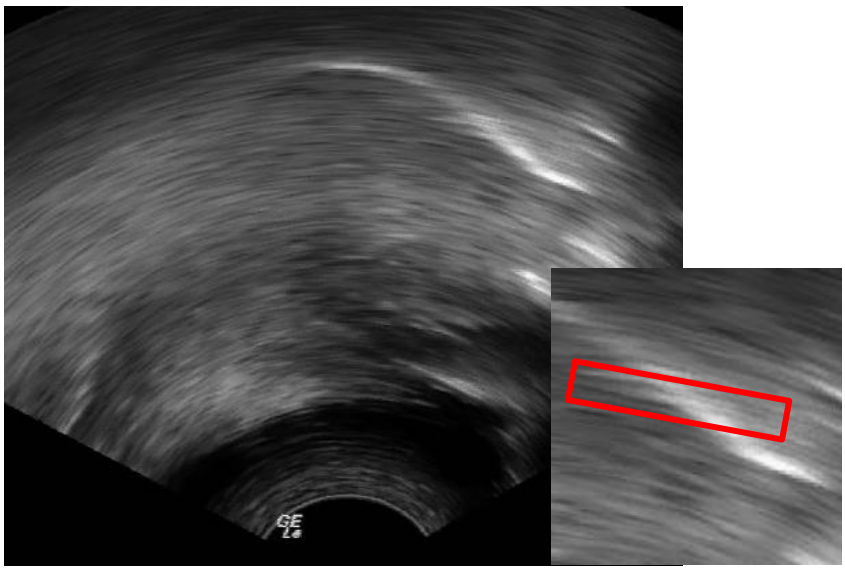


Figure 4. In this image from a GE logiq E running at 124 frames per second it is possible to see the angular smearing due to a combination of interpolation and beam spread.

4 Ultrasonix RP angular (circumferential) resolution

The scanlines are emitted radially and equally spaced. In this case there are 63 spread over 135°. This means there is a gap G of 2.14° between the central axis of each of these beams. At a depth of 6cm from microconvex probe surface (roughly where the adult tongue surface might lie), with a virtual probe radius of 1cm, the angular separation will be calculated by

$$2\pi r(G/360)$$

i.e.

$$2\pi\times7\times(2.14/360) = 2.6\text{mm}$$

Doubling the distance from the virtual probe centre to the tongue surface halves the angular resolution. The resolution as a ratio of the radial distance to the tongue surface is constant, and with 63 scanlines over a 135° FOV, it is approximately 3.7% of the radial distance to the tongue surface. Figure 5a shows the raw return data from these 63 beams, and Figure 5b the more familiar

This is however not the only angular resolution factor to be taken into account. The beams themselves spread out and any surface caught in this diffuse beam, creating an echo, will be assigned to the nominal direction of the beam as part of a linear series of objects whose characteristics are based on the temporal delay of echoes of varying intensity, and plotted as if they all lie along this scanline vector. For microconvex probes it is particularly difficult to form tight beams. Typically they will spread more than 2 degrees. Strong reflections arising some distance from the nominal beam direction can therefore be picked up on either side of it. So for moderate frame rates of 120 fps and lower, beam spread is the predominant factor determining radial resolution (which can also be observed as radial smearing in the ultrasound image). Again, to be clear, this applies to *all* ultrasound systems and not just the ultrasonix RP described here, and can be observed if a calibrated scanning object called a “phantom” is used.

Using a phantom with the ultrasonix RP, reflections from 0.5mm diameter point targets could be observed 2.4° from the nominal beam angle (Figure 6). Resolution of two equidistant targets at a given angular separation will depend on the relative brightness of these target reflections. A very strong reflective target will swamp a very weak target several degrees distant.

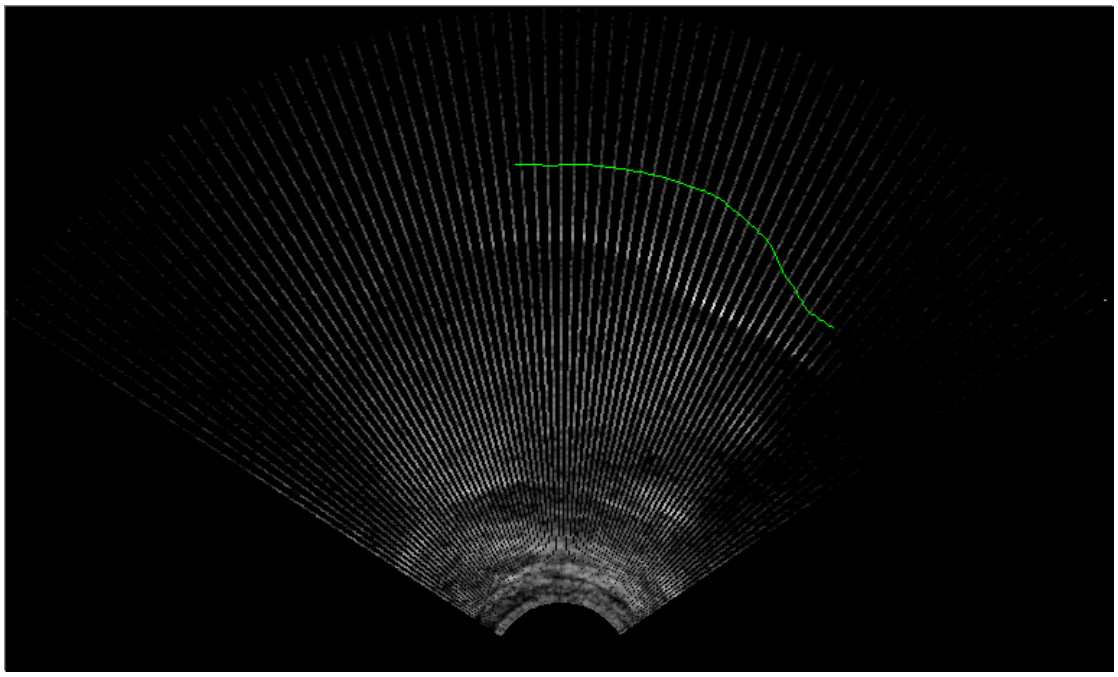


Figure 5a. Shows beams before they are interpolated to form a 733x397 pixel B-mode image, with a constant gap (G) of 2.14°

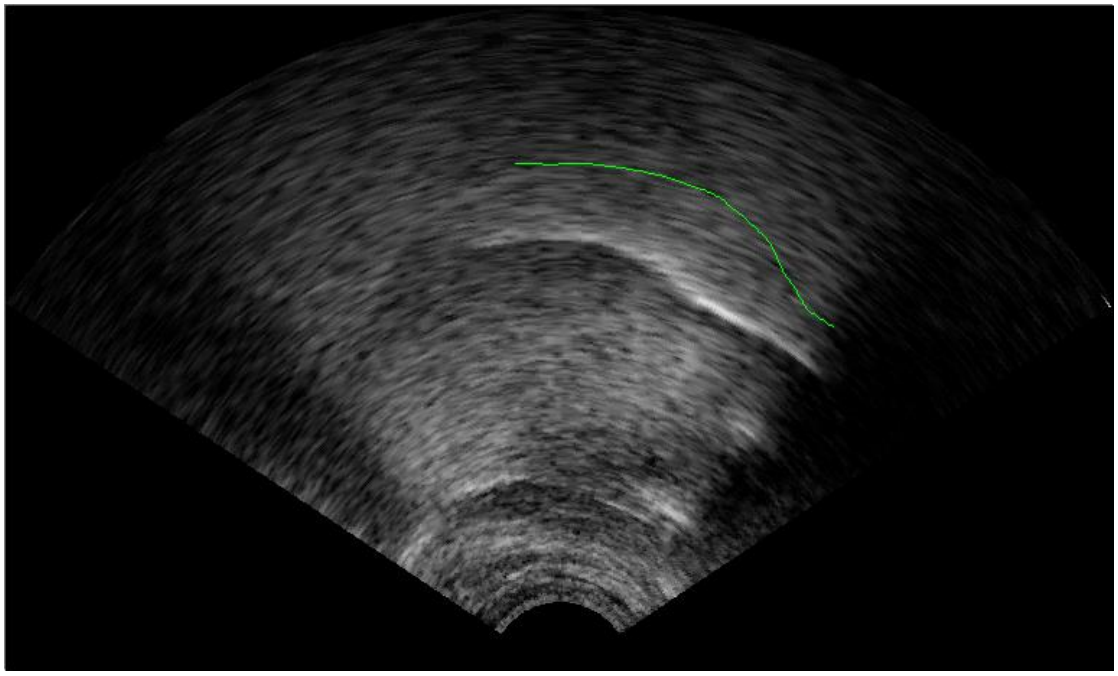


Figure 5b. The raw data in Figure 1 after interpolation to fill the gaps.

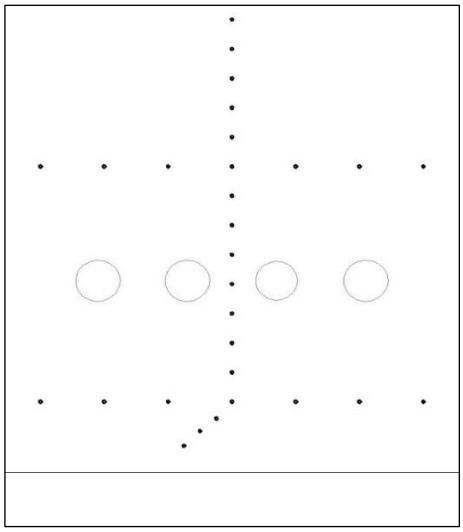
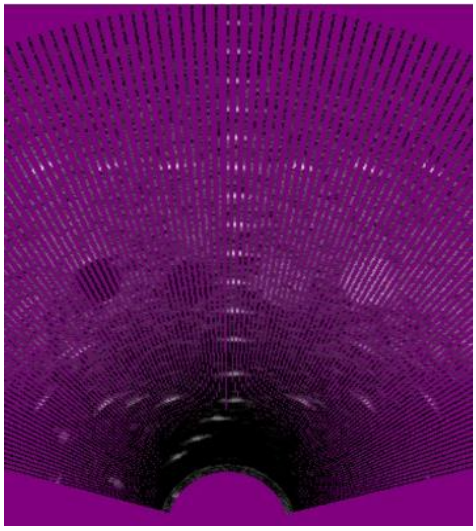
a)  b) 

Figure 6. a) Phantom point target objects b) the echo returns from multiple beams with an angular separation (G) of 1.2° (124 beams over 150° FOV). The point objects at 6cm depth appear on approximately 5 adjacent scanlines i.e a point target is smeared by ±2.4°

5 Axial resolution

So far we have discussed factors affecting angular resolution. The resolution *along* each beam is determined by a different physical property; that of the wavelength of the ultrasound being used. In this paper we use the 5MHz setting on our probe to get maximum penetration. The spatial resolution of ultrasound along the ultrasound beam, is known as the axial resolution (a.k.a. depth, linear, radial, longitudinal or range resolution). The axial resolution is the minimum distance in the beam direction between two reflectors which can be identified as separate echoes. The axial resolution is slightly more than half the spatial pulse length, which is the number of waves in the transmitted ultrasound pulse (determined by the Q factor) multiplied by their wavelength (determined by the transducer frequency). For a typical transducer such as the ultrasonix probe used in the QMU facility the pulse is 3 wavelengths so the axial resolution is ½ x3/(5MHz)x1540m/s = 0.46mm . The digitised echo return data for each beam consists of 412 values leading to a pixel resolution of 8cm/412 = 0.2mm so the pixel resolution is more than twice the axial resolution determined by the probe wavelength. It is, therefore, the latter that limits axial resolution.

6 Alignment accuracy

Alignment accuracy has already been described in the section on synchronisation above but in this short section the alignment will be demonstrated. Ultrasonix RP generates a 3 nanosecond pulse on completion of receiving each beam and on completion of each completed frame.

Figure 7a shows a recording of a metal microphone capsule tapping the centre of the ultrasonic probe in a "tap test". Each ultrasound frame is aligned to the falling edge of the corresponding frame sync pulse. There are 4 ultrasound images. The first (occurring before the tap) is dark; the third and fourth images show a clear "comet tail" image caused by the metal microphone being in contact with the centre section of the probe; the second image shows a partial comet tail image with the lefthand side missing. The tap occurs almost exactly half way through the second frame. At the point in time when the microphone makes initial contact with the probe, all the beams from the left of the arcing sweep to the centre have already been recorded (with no contact and therefore no comet tail). After contact between the microphone and the probe the remaining beams record the comet tail as the scan continues to arc from left to right.

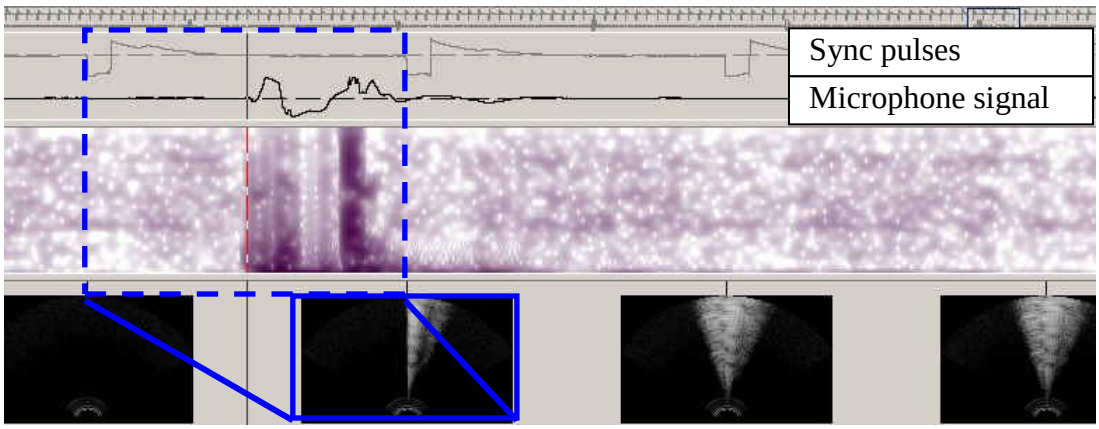


Figure 7 a) Recording of a microphone tapping the Ultrasonix RP probe. The vertical black cursor marks the instant of contact which occurs in this example (by chance) exactly half way through the B-mode scan (frame). The corresponding image shows the flash ("comet trail") from the metal contacting the probe head, which is only visible from individual beams formed after the point of contact.

Figure 7b shows another tap later in the same recording. This time the contact is made 5/8 through a frame and so less of the flash is visible. This tap example highlights the fact that the image is not created at a single instant but formed continuously over the whole period of the frame. It provides evidence not only that the frame from the facility is accurately aligned (to within a fraction of a millisecond) but that each part of any ultrasound image corresponds to a different point in time as well as space. Beams drawn on the left of the RP display are farther back in time than those at the right. Note however that if the probe was rotated so that the tongue tip was pointing to the left, the direction of the scan would be reversed and the beams on the right of the display would be farther back in time.

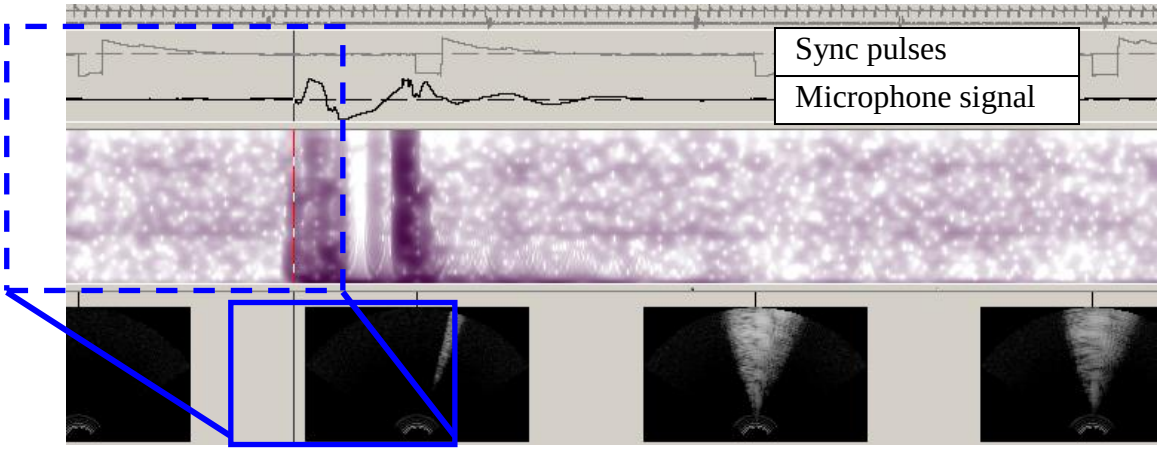


Figure 7 b). Another recording of a microphone tapping the Ultrasonix RP probe. This time the contact occurs 5/8ths of the way through the scan and the image shows the flash only from beams formed after contact is made.

7 Tongue contour estimation

7.1 Fan Grid and manual drawing

The raw ultrasound data captured and stored by AAA is the sequence of returns from each beam/ scanline, grouped into frames: a raw vector set. For convenience of display, these are presented visually as traditional rectangular images with a fan-shaped field of view, with the image filling the space thanks to image interpolation. Unlike most systems, this is undertaken by the AAA software rather than the ultrasound scanner software. Avoiding scanner interpolation as well as having access to well-synchronised ultrasound and audio signals (rather than having access only to the artefact-rich NTSC output of video ultrasound scanners) are important aspects of the QMU facility.

A scaling factor from pixels to mm is automatically calculated by the AAA software based on known recording parameters. For example, if there are 412 pixels per scanline where the scanline covers a depth of 80mm, each pixel represents a nominal . The software allows the user to superimpose a “Fan grid” on the ultrasound image 0.2mm on the display. A fan-shaped grid comprising 42 fan lines is used for analysis, as shown in figure 8. This fan grid is arbitrary and fixed at 42 lines and is entirely unrelated the underlying number of beams in the raw ultrasonic data. 42 was considered a sufficient number to define the subtleties of the shape of the tongue contour. A higher number was felt to be overly sensitive to spurious noise in the image.

Tongue contours can be drawn on this grid defined by a single crossing point on each of the 42 fan lines. So that the contour appears smooth, the connecting line drawn between these 42 points is defined by a B-spline with control points at these crossing points. Manual edge-tracking can be done using a mouse, a pen-pad.

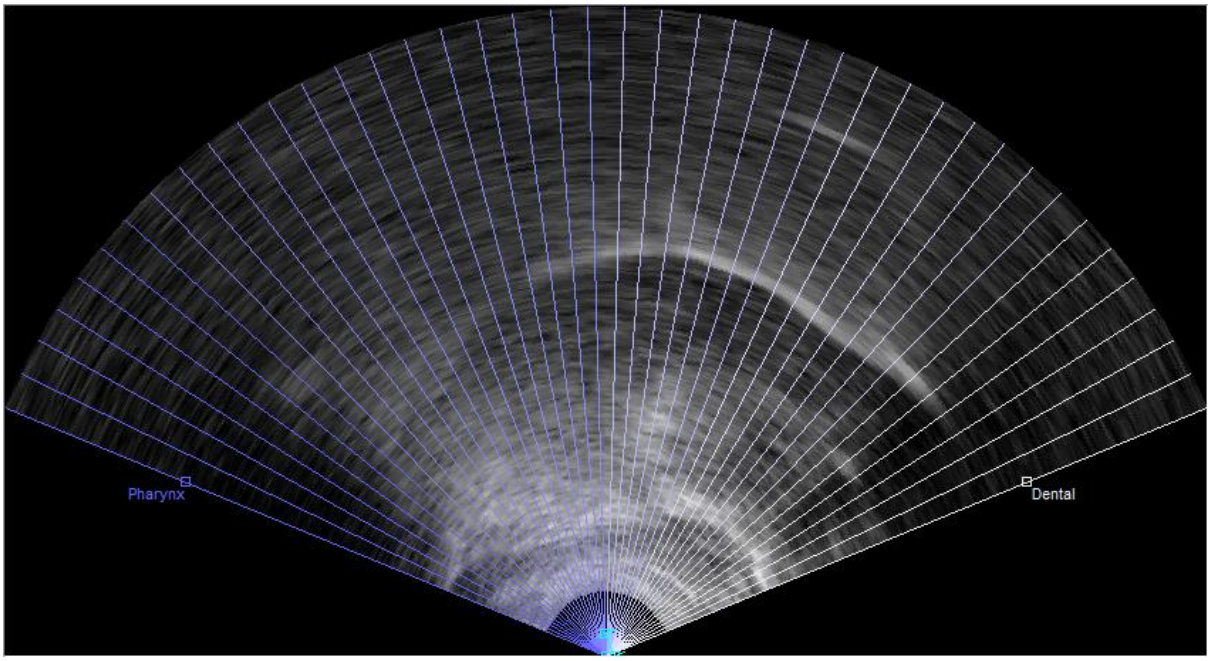


Figure 8 Ultrasound image constructed from raw ultrasonix data with a 42 line fan grid for analysis superimposed.

7.2 Semi-automatic edge detection

Hand fitting contours is time consuming and they are not necessarily appropriately smooth or accurate. AAA software provides a semi-automatic facility to fit tongue contour splines. This method can be applied to the interpolated image rather than the underlying raw data and so can be applied to ultrasound image sequence data imported from any ultrasound system, including the video signal output of standard NTSC systems. Indeed, such data is used in research at QMU as well as elsewhere.

The edge detection algorithm requires a region of interest to be defined by an upper contour limit ("Roof") and lower contour limit ("MinTongue"). These limits, along with a typical tongue contour shape for a segment type of interest, perhaps can be defined in a template and subsequently applied to any new recording as a suitable starting point for edge- fitting.

7.3 The edge detection algorithm

For every 2nd pixel along one of the 42 axes, a triangular weighting function is applied to h=20 pixels on either side (closer and further away from the origin).
An Edge confidence is calculated as C_v= pixel brightness squared multiplied by the weighting function

$$C_v = \sum_{j=-h}^{h} (sign(j)) \frac{pixel^2}{765} \text{(h-abs(j/2))/(h)}$$

The confidence is similarly calculated for 2 further extra radial vectors lying in between the principal axis (v) and its neighbours. These are given less weighting than the principal axis. Figure 9 shows how the weightings are applied.

The highest confidence for each of the 42 fan lines then becomes the basis for the detected edge. A simple moving average smoother is applied to every three adjacent fan lines to provide a smoother contour.

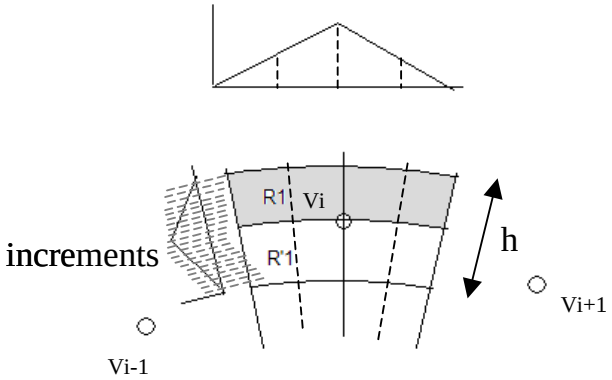


Figure 9 Edge detection confidence weighting.

7.4 Tracking

Once an edge is detected for one frame of an image sequence, "tracking" can be applied for dynamic analysis. Tracking is efficient with ultrasonix RP data (at a high framerate) and a second of data (120 frames) can be tracked in a few seconds. The AAA software assumes that a new contour to be edge-fitted is within 10% of the last known position. This greatly reduces the search area. Such tracking can be applied in forward or backward time direction.

7.5 Confidence

Confidence is a measure of "goodness" of the edge and is defined in the section on edge detection above. However, as the algorithm operates on the image and the image brightness and contrast is a user controlled, there is also a "fiddle factor" that is user defined to scale the confidence to lie in the range 1 to 100. The user typically assigns this value so that where they can see a clear edge in the ultrasound image, the confidence is in the range 80-100 and where there is no edge visible the confidence is in the range 1-20. This value is typically set once and never changed. This enables a user to discard data within a session which fails to exceed a user-defined threshold, on a consistent basis.

The tongue contour is drawn such that regions of the contour with confidence in the range 80-100 form a solid line, 60-80 a dashed line 20-60 a dotted line and 0-20 are invisible.

Confidence can also be used when exporting the x/y co-ordinates of the tongue contour, to exclude any points that fall below a user defined confidence threshold.

7.6 Performance

We have compared the performance of the AAA tracker against EdgeTrak, the popular and freely available "snakes" based spline-fitting algorithm. The performance is similar. The AAA tracker seems not to drift away from the edge as it proceeds as often as EdgeTrak but a rigorous test of this has not been carried out. The fact that the ultrasonix data is recorded at 121Hz means that the frame to frame contour movement is less than on a typical 30 or 60 frames per second ultrasound system and this seems to have a beneficial effect on the tracker performance. In addition, each frame when captured at 121Hz is less smeared: the edge in each frame is crisper.

The tracker, being integrated within the AAA software, is very convenient and can be applied to labelled regions of various sizes. Such labelling can be done within AAA or imported after force-alignment in external packages. One very efficient unit for tracking seems to be the gestural stroke, e.g. the coherent single movement from target to target, e.g. from a vowel target to a following consonant target.

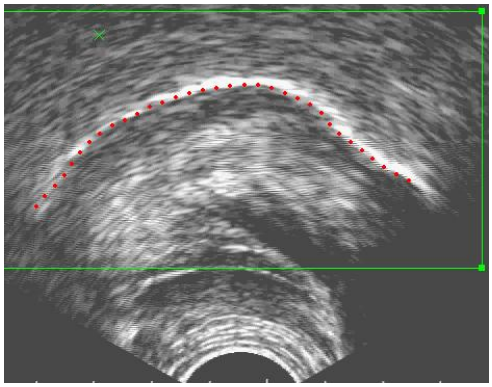
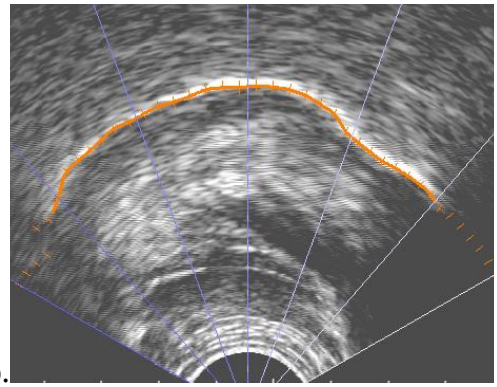
a.  b. 

Figure 10: Comparison of EdgeTrak (left image) with the AAA tracker (right image). Green box denotes region of interest defined for the Edgetrak example.